Coarse-grained protein-protein stiffnesses and dynamics from all-atom simulations

Stephen D. Hicks* and C. L. Henley†
Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, NY 14853-2501

Large protein assemblies, such as virus capsids, may be coarse-grained as a set of rigid domains linked by generalized (rotational and stretching) harmonic springs. We present a method to obtain the elastic parameters and overdamped dynamics for these springs from all-atom molecular dynamics simulations of one pair of domains at a time. The computed relaxation times of this pair give a consistency check for the simulation, and (using a fluctuation-dissipation relationship) we find the corrective force needed to null systematic drifts. As a first application we predict the stiffness of an HIV capsid layer and the relaxation time for its breathing mode.

PACS numbers: 87.10.Pq,87.15.ap,87.15.hg,87.15.Ya

Large protein assemblies—our particular interest is the capsids (shells) of viruses—are pertinent to most of the soft-matter physics in cells; how can one calculate their elastic properties and corresponding dynamics? Such assemblies are too large to handle by all-atom simulations, but numerical coarse graining techniques are opening the door to direct simulations [1]. Nevertheless, we still prefer simplified parametrizations for the purposes of human understanding, analytic treatment, transmission to other researchers, and building up coarse-grained models [2]. In this paper we propose an approach to extract these simplified parameters from all-atom molecular dynamics (MD) of small subsystems.

Assume we already know how to extract an appropriate subsystem and an intelligent way to project the $3N$ coordinates of its configurations onto s small number $\boldsymbol{x}$. (Below, we do this explicitly for the case of the HIV capsid protein.) Our aim is, from the observed trajectories $\boldsymbol{x}(t)$, to extract the parameters for an effective Hamiltonian and equation of motion. We will model $\boldsymbol{x}(t)$ as an overdamped random walk in a biased harmonic potential. This walk is parametrized primarily by two important tensors: one to describe the shape of the harmonic well, and the other to describe the (mainly hydrodynamic) damping and the associated stochastic noise. Combining these tensors gives a matrix whose eigenvalues are the relaxation rates. With detailed measurement of the dynamics, we can identify whether the simulation is equilibrated during the simulation time, and can compute the external forces we must add so as to measure the behavior near the biologically proper configuration. This is similar in spirit to computing a potential of mean force or free energy landscape with Jarzynski's equality[3], except that our coarse-grained $\boldsymbol{x}$ has more than one component, and (at minimum) represents angular degrees of freedom in addition to stretching. As an application, we simulate the important inter-domain interactions in the HIV capsid and estimate the Young's modulus and Poisson ratio of the capsid lattice, as well as the relaxation rate of the breathing mode.

Coarse graining as stochastic dynamics. We represent our system as a vector of generalized coordinates x_i, $i = 1 \ldots N$, where N is far smaller than the number of atoms and is obtained by some form of coarse-graining. Our objective is to parametrize and determine from simulation (i) an effective free energy potential function $U(\boldsymbol{x})$, and (ii) an equation of motion, for the coarse-grained coordinates.

We assume the coarse-grained degrees of freedom are overdamped: this is true at time scales much longer than the "ballistic scale" of local bond vibrations ($t_{\rm bal} \sim 1{\rm ps}$). Then the dynamics is a continuous-time random walk:

$$\frac{\mathrm{d}\boldsymbol{x}}{\mathrm{d}t} = \boldsymbol{\Gamma}\boldsymbol{f}(\boldsymbol{x},t) + \boldsymbol{\zeta}(t), \tag{1}$$

where $\boldsymbol{\Gamma}$ is the (symmetric) *mobility tensor*, $\boldsymbol{f}(\boldsymbol{x},t)$ is the force, $\boldsymbol{\zeta}(t)$ is a (Gaussian) stochastic function satisfying

$$\langle \boldsymbol{\zeta}(t) \otimes \boldsymbol{\zeta}(t') \rangle = 2\mathbf{D}\delta(t-t'), \tag{2}$$

and $\mathbf{D}$ is the *diffusion tensor*. For detailed balance, $\mathbf{D} = k_{\rm B}T\boldsymbol{\Gamma}$ at temperature T. We can expand the potential to second order about a point $\boldsymbol{x}_*$,

$$U(\boldsymbol{x}) = U_0 - \boldsymbol{f}_* \cdot (\boldsymbol{x}-\boldsymbol{x}_*) + \frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_*)\mathbf{K}(\boldsymbol{x}-\boldsymbol{x}_*), \tag{3}$$

where $\mathbf{K}$ is the (symmetric) *stiffness tensor*; then the force in (1) is $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{f}_* - \mathbf{K}(\boldsymbol{x}-\boldsymbol{x}_*)$. From measuring coordinate covariances in the simulation, we obtain $\mathbf{K}$:

$$\mathbf{G} \equiv \big\langle [\boldsymbol{x}-\boldsymbol{x}_*] \otimes [\boldsymbol{x}-\boldsymbol{x}_*] \big\rangle = k_{\rm B}T\mathbf{K}^{-1}. \tag{4}$$

If the static effective potential were our only interest, and if our runs were always long enough to equilibrate our system, there would have been no need to model the dynamics (1). As we do need the dynamics, we determine the diffusion tensor $\mathbf{D}$ (and hence $\boldsymbol{\Gamma}$) by measuring the correlation function at short times between the ballistic and relaxation times scales (see below) during which the deterministic term in (1) is less important than the noise:

$$\mathbf{D} = \frac{\big\langle [\boldsymbol{x}(t')-\boldsymbol{x}(t)] \otimes [\boldsymbol{x}(t')-\boldsymbol{x}(t)] \big\rangle}{2|t'-t|} \equiv \frac{\mathbf{W}(t'-t)}{2|t'-t|}. \tag{5}$$

We average $\mathbf{W}(\Delta t)/|\Delta t|$ over possible offsets $\Delta t \gg t_{\rm bal}$, inversely weighted by the expected variances $\propto (\Delta t)^3$. This weighting also ensures our estimate has negligible contribution from t comparable to the relaxation times, at which times $\mathbf{W}(t)$ is no longer linear in t. Notice that since $\boldsymbol{\Gamma}$ pertains to

short-time dynamics, it is correctly measured even in runs too short to equilibrate in the potential well.

If we transform into coordinates $\widetilde{\boldsymbol{x}} \equiv \boldsymbol{\Gamma}^{-½}\boldsymbol{x}$ then the equation of motion becomes

$$\frac{\mathrm{d}\widetilde{\boldsymbol{x}}}{\mathrm{d}t} = \boldsymbol{\Gamma}^{½}\boldsymbol{f}_* - \mathbf{R}\big(\widetilde{\boldsymbol{x}} - \widetilde{\boldsymbol{x}}_*\big) + \widetilde{\boldsymbol{\zeta}}(t), \tag{6}$$

where

$$\big\langle \widetilde{\zeta}_\alpha(t)\widetilde{\zeta}_\beta(t')\big\rangle = 2k_{\mathrm{B}}T\delta_{\alpha\beta}\delta(t-t'), \tag{7}$$

and the *relaxation matrix* $\mathbf{R} = \boldsymbol{\Gamma}^{½}\mathbf{K}\boldsymbol{\Gamma}^{½}$ (which has units [time]$^{-1}$) is simply the stiffness tensor in our transformed frame. The eigenvalues of $\mathbf{R}$ are the decay rates τ_α^{-1} for the relaxation normal modes α.

The correlation time for a mode is the same as its relaxation time, so the relative error in $\mathbf{K}$ for mode α is of order $\sqrt{\tau_\alpha/\tau_{\text{run}}}$, where τ_{run} is the total run time. Thus, if all the $\tau_\alpha \ll \tau_{\text{run}}$, our estimate (4) of $\mathbf{K}$ is valid. But if $\tau_\alpha \sim \tau_{\text{run}}$ for some direction, not only are errors large, but the initial deviation may still be relaxing over the entire run, which is often visible as a steady drift of the coordinates with mean velocity $\overline{\boldsymbol{v}}$. Averaging over time gives a large spurious variance in the drifting directions, leading to an underestimate of the corresponding stiffness.

Application to HIV capsid. The elastic and dynamic properties of viruses in general are of particular importance in understanding the mechanisms by which they assemble and disassemble. The assembly must be reliable enough to produce capsids capable of surviving the harsh intercellular environment, while still being able to disassemble upon entering a new host cell. HIV in particular is unique because of its characteristic conical capsids [4], whose mechanism of formation is yet unsettled.

A capsid is well-modeled by a triangular lattice of proteins, and we coarse-grain at this level. We take rigid units to represent either a whole protein or a sub-domain of a protein. Each unit therefore requires six coordinates for its position and orientation. Provided the actual interactions are pairwise between units, our program is to *simulate only a pair of interacting units* at a time, doing a separate simulation for each kind of unit-unit contact to obtain its parameters. The coarse-grained network is then reassembled and studied using these generalized springs.

The HIV capsid protein (CA) consists of two globular domains: the larger 145-amino acid N-terminal domain (NTD) has a radius $1.3\mathrm{nm}$ and the smaller 70-amino acid C-terminal domain (CTD) has a radius $1.7\mathrm{nm}$; we treat these as two separate units. The NTD and CTD are connected covalently by a flexible linker; there is also an NTD–NTD interaction (which forms hexamers in the capsid structure), a CTD–CTD interaction (which forms symmetric dimers in the structure), and an NTD–CTD interaction between neighboring proteins around a hexamer. These four interactions are shown in FIG. 1(a). We believe the NTD–CTD interaction to be the weakest, and the known structure is also poorest, so we will ignore it from now on. We therefore simulate each other pair in isolation, using structures from the Protein Data Bank [5].

We carried out our simulations using a modified version of the NAMD [6] package with the CHARMM22 force field. Our proteins are in a periodic cell 5 to 9nm to a side using the TIP3P model for explicit water and 0.1M NaCl, run with 2fs timesteps for a total of 3ns each. We do most of the work in at constant pressure and temperature (NPT), using a Langevin piston barostat at $P = 1\mathrm{atm}$, and a Langevin thermostat at $T = 310\mathrm{K}$ and damping rate $\gamma_{\mathrm{L}} = 5\mathrm{ps}^{-1}$. The NPT simulations model the statics well, but the thermostat's damping leads to unphysical dynamics with increased relaxation rates. This allows shorter simulations to equilibrate, but prevents us from determining the rates we should expect to see in reality. We therefore do a second measurement of diffusion at constant volume and energy (NVE).

The center of mass and global rotation of the pair accounts for six trivial degrees of freedom; the remaining six represent the relative position and orientation of the two domains. Of these six, only one is a pure translation: the distance $r = |\boldsymbol{r}_2 - \boldsymbol{r}_1|$ between the center of each domain. The orientation of domain m can be represented by a rotation matrix $\boldsymbol{\Omega}_m$ which rotates the domain from its reference orientation by an angle $|\boldsymbol{\theta}_m|$ about the axis $\hat{\boldsymbol{\theta}}_m$. The even and odd combinations $\boldsymbol{\theta}_1 \pm \boldsymbol{\theta}_2$ give six degrees of freedom that comprise the remaining five coordinates, along with an overall rotation due to the even combination about the inter-body axis $\boldsymbol{r}_2 - \boldsymbol{r}_1$.

As we simulate just one pair of units from a protein complex, we omit the forces and torques on them due to the other units in the lattice, which generically had a nonzero resultant. In order to expand the free energy around the physiologically relevant configuration, we must add external forces to compensate; in light of (1) the correct force to impose is given by $\boldsymbol{f}_* = -\boldsymbol{\Gamma}^{-1}\overline{\boldsymbol{v}}$, where $\overline{\boldsymbol{v}}$ is the drift velocity measured in the absence of the compensating force. This was not important for the pairs reported in our results.

Results. The results for each simulation were similar, and the trajectory of the linker in the transformed relaxation mode coordinates is shown in FIG. 1(b), which is characteristic of all the observed trajectories. Once we have an equilibrated segment of a trajectory we use (4) to determine the 6×6 stiffness tensor $\mathbf{K}$; different components have different units, so it would be mathematically meaningless to diagonalize it directly. Instead, we define reduced stiffness tensors, representing the free energy cost if we optimize r for a fixed set of angles and vice versa. Given

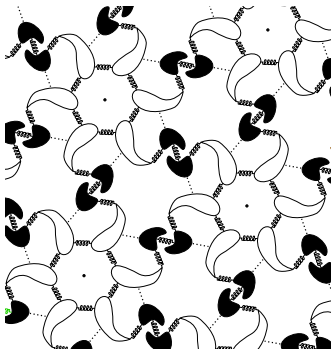
$$\mathbf{K} = \begin{pmatrix} K_{\mathrm{rr}} & \mathbf{K}_{\mathrm{r}\theta} \\ \mathbf{K}_{\theta\mathrm{r}} & \mathbf{K}_{\theta\theta} \end{pmatrix}, \tag{8}$$

then

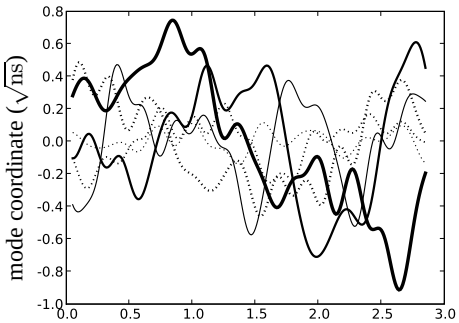
$$K^{(\mathrm{eff})}_{\mathrm{stretch}} = K_{\mathrm{rr}} - \mathbf{K}_{\mathrm{r}\theta}\mathbf{K}^{-1}_{\theta\theta}\mathbf{K}_{\theta\mathrm{r}} \tag{9}$$

$$\mathbf{K}^{(\mathrm{eff})}_{\mathrm{orient}} = \mathbf{K}_{\theta\theta} - \mathbf{K}_{\theta\mathrm{r}}K^{-1}_{\mathrm{rr}}\mathbf{K}_{\mathrm{r}\theta}. \tag{10}$$

The eigenvalues of the reduced tensors are given in TABLE I.



(a)



(b)

FIG. 1: (a) Diagram of interactions in the HIV capsid lattice. The black and white shapes represent the dimer-forming CTD and the hexamer-forming NTD, respectively. Springs represent the three different bonds we are interested in, and dotted lines represent the fourth bond we are ignoring. (b) Relaxation mode trajectories of linker. The mode coordinate has units of $\sqrt{\text{ns}}$ because it has been normalized by the noise. The slower modes are drawn with thicker lines. Note that the slowest mode has a very small drift, and we could correct this by applying an external force. The traces have been smoothed with a low-pass filter for readability.

	$K^{(\text{eff})}_{\text{stretch}}$ ($k_\text{B}T/\text{nm}^2$)	$\mathbf{K}^{(\text{eff})}_{\text{orient}}$ eigenvalues ($k_\text{B}T$)				
NTD–NTD	12	1300	2800	4500	10000	18000
CTD–CTD	9.9	210	340	1100	3900	8300
Linker	2.8	130	250	480	1100	3800

TABLE I: Effective stiffness eigenvalues for pair simulations: NTD dimer, CTD dimer, and the NTD–CTD linker within the CA protein.

	relaxation times τ_α (ps)					
NTD–NTD	120	23	18	9.3	6.0	4.4
CTD–CTD	76	26	24	7.8	5.4	4.1
Linker	190	140	80	76	22	8.3

TABLE II: NPT time constants for the relaxation modes of each pair.

We computed the stiffness tensor implicitly in the relative coordinates between the two bodies, but the absolute coordinates are the natural frame for computing the noise. Measuring the diffusion of a single body in an NVE simulation yields a mean $D^{(\text{rot})}_{\text{CTD}} = 0.11\text{rad}^2/\text{ns}$ and $D^{(\text{rot})}_{\text{NTD}} = 0.044\text{rad}^2/\text{ns}$. If we approximate each domain as a solid sphere then Stokes' law gives a rotational diffusion constant $D^{(\text{rot})} = k_\text{B}T/(8\pi\eta r^3)$ [7]. We thus expect $D^{(\text{rot})}_{\text{CTD}} = 0.11\text{rad}^2/\text{ns}$ and $D^{(\text{rot})}_{\text{NTD}} = 0.050\text{rad}^2/\text{ns}$ using a viscosity $\eta^{(310\text{K})} = 0.69\text{cP}$. The accepted TIP3P viscosity $\eta^{(\text{TIP3P})} = 0.31\text{cP}$ gives poorer agreement.

The translational diffusion constant is slightly harder to measure, since it is influenced significantly by the finite-size effect [8]. This can be corrected for by measuring the diffusion at several box side lengths L and using a linear fit of $D^{(\text{tr})}$ versus $1/L$ to extrapolate to $1/L = 0$. Doing so yields $D^{(\text{tr})}_{\text{CTD}} = 55\text{Å}^2/\text{ns}$ and $D^{(\text{tr})}_{\text{NTD}} = 27\text{Å}^2/\text{ns}$. Stokes' law gives expected $D^{(\text{tr})}_{\text{CTD}} = 56\text{Å}^2/\text{ns}$ and $D^{(\text{tr})}_{\text{NTD}} = 43\text{Å}^2/\text{ns}$ using $\eta^{(\text{TIP3P})} = 0.31\text{cP}$. The measured $D^{(\text{tr})}$ has a significantly larger relative error than $D^{(\text{rot})}$, due to the finite-L extrapolation.

We can diagonalize the relaxation matrix to compute the relaxation modes for each linkage. The NPT relaxation times from this calculation are listed in TABLE II. All the times are significantly shorter than the simulation time, so we can be confident that the simulations are equilibrated.

Finally, we can compose these generalized springs together into a triangular lattice as shown in FIG. 1(a), with an NTD hexamer at each vertex, a CTD dimer at the midpoint of each edge, and a spring connecting each domain, whose free energy is given by the relative positions multiplied into the appropriate stiffness tensor. We can then determine the free energy minimum as a function of periodic cell dimensions to find a lattice constant of $a = 9.1\text{nm}$. This is slightly smaller than the experimentally measured 10.7nm [4], which may be largely due to our sheet being flat, rather than curved into a tube. Computing the free energy of simple extension yields a 2d Young's modulus of $0.92k_\text{B}T/\text{Å}^2 = 0.39\text{N/m}$ and a Poisson ratio of 0.30. Assuming homogeneity and a thickness of 5nm, we find a 3d Young's modulus of 77MPa (compared with 115MPa measured using atomic force microscopy [9]).

Furthermore, we can estimate the relaxation rate of the full-capsid breathing mode in water by further coarse-graining to a single coordinate a representing a uniform dilation in the plane, which has dynamics given by (1) with stiffness and mobility constants K_a and Γ_a. The projected stiffness is given by the bulk modulus $K_a = 4K_{2\text{d}} = 2.6k_\text{B}T/\text{Å}^2$, calculated from the 2d Young's modulus and Poisson ratio. To project the damping term, we observe that all the actual motion in the breathing mode of a virus capsid of radius r is in the radial direction, and we thus need to scale the capsid protein's translational diffusion constant by $(\text{d}a/\text{d}r)^2$ to find the diffusion

constant for a. Using the detailed balance condition,

$$\Gamma_a = \frac{16\pi\sqrt{3}}{N} \frac{D_{\rm NTD}^{\rm (tr)} + D_{\rm CTD}^{\rm (tr)}}{k_{\rm B}T} \frac{\eta^{\rm (TIP3P)}}{\eta^{\rm (310K)}}, \qquad (11)$$

where $N = 16\pi\sqrt{3}r^2/a^2$ is the total number of capsid proteins [10]. Taking $N = 1500$ proteins as the average size for an HIV capsid thus gives a relaxation rate of $6.1{\rm ns}^{-1}$ for the breathing mode.

Discussion. In conclusion, we have put forth a model of overdamped random walks in which the statics and dynamics are described respectively by complementary "stiffness" and "mobility" tensors. From these two tensors a "relaxation matrix" can be formed, the eigenvalues of which give the relaxation rates, which also provide a convergence test for simulations. We demonstrated the usefulness of this model in extracting coarse-grained elastic constants from molecular dynamics trajectories of pairs of interacting domains. HIV is particularly well-suited for this because the important interactions appear to be nearest-neighbor, while many other viruses have long tails in which all six molecules in the hexamer are entwined, making it more difficult to separate into individual interactions.

Evaluating the forces between protein domains to second order in the positions and orientations yields a picture of the dynamics that is simple enough both to simulate with all-atom MD as well as to model at the coarse-grained level, yet general enough to thoroughly describe the interaction in the vicinity of the simulated configuration.

Our relaxation formalism bears some similarities to normal mode analysis, and in particular, Gaussian network models, which replace atomic interactions by springs of uniform stiffness [11]. While these techniques have been successful in explaining reaction pathways such as virus maturation [12, 13], they suffer from several shortcomings: first, while the normal mode frequencies are useful in identifying soft degrees of freedom, the frequencies themselves are well known to be artificial because they omit the damping forces of the surrounding water. For instance, the breathing mode we computed above would have a normal mode frequency of $\sqrt{K/m} = 60{\rm ns}^{-1}$. Additionally, most applications are coarse-grained to the point that individual residue types are irrelevant: such a method is entirely insensitive to the effect of point mutations or of varying the salinity. Lamm and Szabo [14] introduced so-called "Langevin modes," which are similar to our relaxation modes, but their method still suffers from the latter issues.

Another quantitative approach to understanding protein dynamics is "essential dynamics" (or "principle component analysis") [15]. This technique has the advantage that it is based on all-atom simulations, with the explicit damping forces and entropic contributions of the solvent, but the resulting modes can only be expressed by giving a $3N$-component vector. Hayward *et al* [16] suggested specifying important modes *a priori*, and this provides us the great advantage being able to relate the results of several simulations together into a bigger picture. As long as our modes still contain the most important fluctuations, they are a reasonable basis to use.

We have demonstrated the use of relaxational dynamics in extracting measurable elastic moduli from small molecular dynamics simulations. We hope that this technique will provide a convenient middle ground between the atomistic and continuum pictures for other biological systems.

Acknowledgments. We thank D. Murray, V. M. Vogt, M. Widom, H. Weinstein, D. Roundy, W. Sundquist, M. Yeager, and P. Freddolino. This work was supported by DOE Grant No. DE-FG02-89ER-45405. Computing facilities were provided through the Cornell Center for Materials Research under NSF grant DMR-0079992.

* Electronic address: sdh33@cornell.edu
† Electronic address: clh@ccmr.cornell.edu

[1] A. Arkhipov, P. Freddolino, and K. Schulten, Structure **14**, 1767 (2006).
[2] R. Zandi *et al*., Proc. Nat. Acad. Sci. U.S.A. **101**, 15556 (2004); S. D. Hicks and C. L. Henley, Phys. Rev. E **74**, 031912 (2006); M. F. Hagan and D. Chandler, Biophys. J. **91**, 42 (2006).
[3] C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997); Eur. Phys. J. B **64**, 331 (2008).
[4] S. Li, C. P. Hill, W. I. Sundquist, and J. T. Finch, Nature **407**, 409 (2000).
[5] For the full-length protein (linker simulation) we use cryo-EM structure 3DIK (B. Ganser-Pornillos, A. Cheng, and M. Yeager, Cell **131**, 70 (2007)); for the NTD we use the NMR structure 1GWP (C. Tang, Y. Ndassa, and M. Summers, Nat. Struct. Biol **9**, 537 (2002)) fitted to the homologous MLV hexamer crystal structure 1U7K (G. Mortuza *et al*, Nature **431**, 481 (2004)); and for the CTD we use the crystal structure 1AUM (T. Gamble *et al*, Science **278**, 849 (1997)).
[6] J. Phillips *et al*, J. Comput. Chem. **26**, 1781 (2005).
[7] H. Lamb, *Hydrodynamics* (Cambridge, 1932), 6th ed., p. 589.
[8] I. C. Yeh and G. Hummer, Biophys. J. **86**, 681 (2004).
[9] N. Kol *et al*., Biophys. J. **92**, 1777 (2007).
[10] We have included a term $\eta^{\rm (TIP3P)}/\eta^{\rm (310K)}$ to correct for the TIP3P water model's incorrect viscosity (which affected the diffusion constants we measured in our simulations) so that we can estimate the relaxation rate in real water.
[11] M. M. Tirion, Phys. Rev. Lett. **77**, 1905 (1996); I. Bahar, A. R. Atilgan, M. C. Demirel, and B. Erman, Phys. Rev. Lett. **80**, 2733 (1998); F. Tama, M. Valle, J. Frank, and C. L. Brooks, Proc. Nat. Acad. Sci. U.S.A. **100**, 9319 (2003); M. Gibbons and W. Klug, J. Mat. Sci. **42**, 8995 (2007).
[12] A. Rader, D. Vlad, and I. Bahar, Structure **13**, 413 (2005).
[13] E. R. May (personal communication) has calibrated the stiffnesses in an elastic network model to all-atom MD of the HK97 phage (mature) capsid.
[14] G. Lamm and A. Szabo, J. Chem. Phys. **85**, 7334 (1986).
[15] T. Horiuchi and N. Go, Proteins **10**, 106 (1991); T. Ichiye and M. Karplus, Proteins **11**, 205 (1991); A. Amadei, A. Linssen, and H. Berendsen, Proteins **17**, 412 (1993).
[16] S. Hayward, A. Kitao, and H. Berendsen, Proteins **27**, 425 (1997).